Linear regression model selection using p-values when the model dimension grows

By Piotr Pokarowski [1], Jan Mielniczuk [2] and PawełTeisseyre [3]

Abstract. We consider a new criterion-based approach to model selection in linear regression. Properties of selection criteria based on p-values of a likelihood ratio statistic are studied for families of linear regression models. We prove that such procedures are consistent i.e. the minimal true model is chosen with probability tending to 1 even when the number of models under consideration slowly increases with a sample size. The simulation study indicates that introduced methods perform promisingly when compared with Akaike and Bayesian Information Criteria.
Keywords: model selection criterion; random or deterministic design linear model; p-value based methods; Akaike Information Criterion; Bayesian Information Criterion.

1 Introduction

We reconsider a problem of model choice for a linear regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations which variability we would like to explain, $\mathbf{X}$ is a $n \times M_n$ design matrix consisting of vectors of M_n potential regressors collected from n objects and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is an unknown vector of errors, assumed to have $N(0, \sigma^2\mathbf{I})$ distribution. Vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{M_n})'$ is an unknown vector of parameters. In the paper we will consider the cases corresponding to experimental and observational data when rows of $\mathbf{X}$ are either deterministic or random. Suppose that some covariates are unrelated to the prediction of $\mathbf{Y}$, so that the corresponding coefficients β_i are zero. It is assumed that the true model is a submodel of (1). As it is not a priori known which variables are significant in order to make the last assumption realistic it is natural to let the horizon M_n to grow with n and allow in this way potentially large models.

Model selection is a core issue of statistical modeling. In a framework of linear regression the problem has been intensively studied under various conditions imposed on design matrix $\mathbf{X}$ and growth of M_n. The aim of such procedures is to choose the most parsimonious model describing adequately a given data set. For the review of these advances we refer to Pötscher and Leeb (2008). The main problem here is a modeler's dillema that underfitting leads to omission of important variables in the model whereas overfitting involves unnecessary parameter estimation for redundant coefficients which lessens the precision of the model fit.

In the article we contribute to a line of research in which the chosen model is the maximiser of a chosen

[1]Institute of Applied Mathematics and Mechanics, Faculty of Mathematics, Informatics and Mechanics, University of Warsaw. E-mail: pokar@mimuw.edu.pl.
[2]Institute of Computer Science, Polish Academy of Sciences and Faculty of Mathematics and Information Sciences, Warsaw University of Technology. E-mail: miel@mini.pw.edu.pl.
[3]Institute of Computer Science, Polish Academy of Sciences. E-mail: teisseyrep@ipipan.waw.pl.

criterion function. In a seminal paper which is typical for this approach Akaike (1970), starting with the idea of maximising the expectation of predictive likelihood, has shown that the usual likelihood has to be modified to obtain an unbiased estimator of the expectation. The likelihood modified in such a way is known as Akaike Information Criterion (AIC). Variety of other modifications of the likelihood followed, with Bayes Information Criterion (BIC) being the most frequently used competitor. Recently, Pokarowski and Mielniczuk (2010) introduced model selection criteria mPVC and MPVC based on p-values of a likelihood ratio statistic for families of linear models with deterministic covariates and constant dimension. The idea in the case of minimal p-value criterion mPVC is to consider the model selection problem from a point of view of testing a certain null hypothesis H_0 against several hypotheses H_i and to choose the hypothesis (the model) for which the null hypothesis is most strongly rejected in its favour. The decision in the case of mPVC is based on a new criterion which is the minimal p-value of the underlying test statistics. We stress that the discussed selection method is based on a completely different paradigm than the existing approaches: instead of penalizing the likelihood ratio statistic directly by subtracting a complexity penalty its appropriate function is chosen as a selection criterion.

We study conditions under which such a rule is consistent i.e. it choses the minimal true model with probability tending to 1 when the sample size increases. Our main theoretical result stated in Theorem 1 asserts that this property holds for the minimal p-value criterion mPVC provided M_n increases at a slower rate than $\log n + a_n$ where a_n are weights appearing in the scaling of p-values. Similar result is proved for maximal p-value criterion MPVC. Both results apply also to the case when M_n is constant provided the full model (1) is correctly specified. We also introduce and investigate less computationally demanding greedy versions of the discussed methods.

In the last section we present the results of limited simulation study which shows that the introduced methods perform on average better than AIC and BIC criteria. In particular, their performance measured by probability of correct subset detection and prediction error is much more stable when the length of list of models M_n increases i.e. regression model becomes sparse.

In the paper we focus mainly on explanation i.e. finding the model which adequately describes the data. Besides the immediate application of model selection methods to to the second main task of prediction let us mention their use in construction of data-adaptive smooth tests (see e.g. Ledwina (1994)).

Problem of linear model selection when the number of possible predictors increases with the sample size has been studied from different angle by Shao (1997) who defined the optimal submodel to be submodel minimizing the averaged squared prediction error and investigated conditions under which the selected model converges in probability to this model. Moreno et al. (2010) considered Bayesian approach to this problem and proposed using Bayes factors for intrinsic priors as selection criteria.

The main contribution of the present paper is establishing consistency of the criteria based on p-values when the linear model dimension grows. The result is proved for the random design as well as for the fixed design scenario, the former being treated in detail. Intrumental in the proofs are Lemmas 3, 4, 5 which can be also useful for different purposes.

The paper is organized as follows. In Section 2 we introduce considered selection criteria. In Section

3 we discuss the imposed assumptions and consistency results for the family of models consisting of all subsets of predictors as well as hierarchic family. We also introduce greedy modifications of the considered criteria. Section 4 contains proofs of the main results and Section 5 discussion of the results of numerical experiments. Proofs of some auxiliary lemmas are relegated to the Appendix.

2 Model Selection criteria for linear regression models based on p-values

We start by explicitly stating the basic assumption we impose on random-design regression model. Assume that the rows $\mathbf{x}_1', \ldots, \mathbf{x}_n'$ of a matrix $\mathbf{X}(n \times M_n)$ are iid, $\mathbf{x}_l = \mathbf{x}_l^{(n)} = (x_{l,1}^{(n)}, \ldots x_{l,M_n}^{(n)})'$, $l = 1, \ldots, n$. Throughout we consider the situation that the minimal true model is fixed i.e. it does not change with n. Vectors $\{\mathbf{x}_1^{(n)'}, \ldots, \mathbf{x}_n^{(n)'}\}$ constitute rows in an array of iid sequences of M_n-dimensional random variables. We impose the condition that M_n is nondecreasing and that the law of the first M_n coordinates of $\mathbf{x}_1^{(n+1)}$ coincides with that of $\mathbf{x}_1^{(n)}$ i.e. the distribution of attributes considered for a certain sample size remains the same for larger sample sizes. We also assume throughout that the second moments of coordinates of $\mathbf{x}_1^{(n)}$ are finite for any n. As any submodel of (1) containing p_j variables $(x_{l,j_1}^{(n)}, \ldots, x_{l,j_{p_j}}^{(n)})'$ can be described by set of indexes $j = \{j_1, \ldots, j_{p_j}\}$ in order to make notation simpler it will be referred to as model j. The minimal true model will be denoted by t and p_t will be the number of nonzero coefficients in equation (1). The empty model $\mathbf{Y} = \boldsymbol{\varepsilon}$ will be denoted briefly by 0 and the full model (1) by $f = \{1, \ldots, M_n\}$. Note that $M_n = p_f$. Let $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j_1}, \ldots, \hat{\beta}_{j_{p_j}})'$ be a maximum likelihood (ML) estimator of $\boldsymbol{\beta}$ calculated for the considered model j. We denote $\hat{\boldsymbol{\beta}}_f$, ML estimator in the full model, briefly by $\hat{\boldsymbol{\beta}}$. Let $\mathcal{M}$ be a certain family of subsets of a set f and $\mathbf{x}_{lt} = (x_{l,t_1}^{(n)}, \ldots, x_{l,t_{p_t}}^{(n)})'$ be a vector of variables which pertain to the minimal true model t. Througout this paper with exception of Section 3.2 we will impose the following assumption:

(A0) $\mathbf{E}(\mathbf{x}_{1t}\mathbf{x}_{1t}')$ is positive definite matrix.

The main objective of model selection is to identify the minimal true model t using data $(\mathbf{X}, \mathbf{Y})$. Let $\mathbf{f}_{\boldsymbol{\beta},\sigma^2}(\mathbf{Y}|\mathbf{X})$ be the conditional density of $\mathbf{Y}$ given $\mathbf{X}$. Consider two models j and k where the first model is nested within the second model. Denote by D_{jk}^n likelihood ratio test (LRT) statistic, based on conditional densities given $\mathbf{X}$, for testing H_0 : model j is adequate against hypothesis H_1 : model k is adequate whereas j is not, equal to

$$D_{jk}^n = 2 \log \frac{\mathbf{f}_{\hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k^2}(\mathbf{Y}|\mathbf{X})}{\mathbf{f}_{\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2}(\mathbf{Y}|\mathbf{X})}, \tag{2}$$

where $\hat{\sigma}_j^2 = RSS(j)/n$ and $RSS(j)$ is a sum of squared residuals from the ML fit of the model j. We recall that ML estimator $\hat{\boldsymbol{\beta}}_k$ coincides with Least Squares estimator of $\boldsymbol{\beta}$. When j and k are linear models it turns out that LRT statistic is given explicitly by

$$D_{jk}^n = -n \log \left[\frac{RSS(k)}{RSS(j)} \right] = -n \log(1 - R_{jk}^n),$$

where

$$R^n_{jk} = \frac{RSS(j) - RSS(k)}{RSS(j)} \tag{3}$$

is coefficient of partial determination of variables belonging to $k \setminus j$ given that variables in set j are included in the model. Under the null hypothesis H_0 it follows from Cochran's theorem (cf. e.g. Section 5.5 in Rencher and Schaalje (2008)) that given $\mathbf{X}$ $RSS(j) \sim \sigma^2 \chi^2_j$ and $R^n_{jk} \sim Beta(\frac{p_k - p_j}{2}, \frac{n-p_k}{2})$ provided $\mathbf{X}$ is of full column rank.

Let F and G be univariate cumulative distribution functions and T be a test statistic which has distribution function G not necessarily equal to F. Let $p(t|F) = 1 - F(t)$. By p-value of a test statistic T given distribution F (null distribution) we will mean $p(T|F)$. We will consider p-values of statistic R^n_{jk} given Beta distribution with shape parameters $\frac{p_k - p_j}{2}$ and $\frac{n-p_k}{2}$. In order to make notation simpler $p(R^n_{jk}|Beta(\frac{p_k - p_j}{2}, \frac{n-p_k}{2}))$ will be denoted as $p(R^n_{jk}|p_k, p_j)$. We define the following model selection criteria based on p-values of statistic R^n_{jk} when one of the indices is held fixed and the other ranges over all potential models.

Minimal p-value Criterion (mPVC)

$$M^n_m = \text{argmin}_{j \in \mathcal{M}} e^{p_j a_n} p(R^n_{0j}|p_j, 0),$$

where $p(R^n_{00}|0, 0) = e^{a_n}/\sqrt{n}$ and (a_n) is a sequence of nonnegative numbers. When a minimizer is not unique, the set with the smallest number of elements is chosen. In the case of ties, arbitrary minimizer is selected. Observe that when $a_n \equiv 0$ then from among the pairs $\{(H_0, H_j)\}$ we choose a pair for which we are most inclined to reject H_0 and we select the model corresponding to the most convincing alternative hypothesis. For positive a_n the scaling factor $e^{p_j a_n}$ is interpreted as additional penalization for the complexity of a model.

Moreover, Maximal p-value Criterion is defined as

Maximal p-value Criterion (MPVC)

$$M^n_M = \text{argmax}_{j \in \mathcal{M}} e^{-p_j a_n} p(R^n_{jf}|M_n, p_j),$$

where $p(R^n_{ff}|M_n, M_n) = 1$ and $a_n \to \infty$. Thus from among the pairs $\{(H_j, H_1)\}$ we choose a pair for which we are most reluctant to reject H_0 in favour of the full model hypothesis. We stress that the additional assumption $a_n \to \infty$ needed for consistency of MPVC is not required to prove consistency of mPVC. This point is discussed further in Section 3. Note that in the definition of both criteria the existence of encompassing model, either from below or from above, is vital for the construction. The idea of encompassing has been used in Bayesian model selection (see e.g. Casella et al. (2009)).

Observe that for a fixed number of variables p_j p-value $p(R^n_{0j}|p_j, 0)$ is a strictly decreasing function of R^n_{0j}. Thus the set M^n_m is actually chosen from among subsets for which R^n_{0j} is maximal for the stratum $p_j = 1, \ldots, M_n$. The same observation also holds for MPVC as well as for BIC and AIC. Observe also that if these criteria choose subsets of the same cardinality, these subsets necessarily coincide.

3 Results

3.1 Random-design regression

The main result of this section is consistency of the introduced selectors. Depending on the context we will use some of the following additional conditions on the horizons M_n, norming constants a_n and matrix $\mathbf{X}$.

(A1.1') $M_n/(a_n + \log(n)) \to 0$ as $n \to \infty$.

(A1.1") $M_n/a_n \to 0$ as $n \to \infty$.

(A1.2) $\lim_{n\to\infty} M_n \geq \max_{i\in t} i =: i_{\max}$.

(A1.3) The minimal eigenvalue κ_n of $\mathbf{E}[\mathbf{x_1}^{(n)}\mathbf{x_1}^{(n)'}]$ is bounded away from zero, i.e. $\kappa_n > \kappa > 0$ for some $\kappa > 0$ and $n \in \mathbf{N}$.

(A1.4) For some $\eta > 0$, $n^{-1}M_n^{1+\eta} \to 0$ and

$$\sup_n \sup_{||\mathbf{d}||=1} \mathbf{E}|\mathbf{d}'\mathbf{z}^{(n)}|^{4\lceil 2/\eta\rceil} < \infty, \tag{4}$$

where $\mathbf{z}^{(n)} = \mathbf{E}[\mathbf{x}_1^{(n)}\mathbf{x}_1^{(n)'}]^{-1/2}\mathbf{x}_1^{(n)}$ is the standardised vector $\mathbf{x}_1^{(n)}$ i.e. $E(\mathbf{z}^{(n)}\mathbf{z}^{(n)'}) = \mathbf{I}$ and $\lceil 2/\eta\rceil$ is the smallest integer greater than or equal to $2/\eta$.

(A1.5) $a_n/n \to 0$ as $n \to \infty$.

Assumptions (A1.1') and (A1.1") are two variants of the condition on a rate of divergence of M_n. As M_n is nondecreasing, the limit in (A1.2) exists and is either finite or equal to infinity. Condition (A1.2) is a natural condition stating that ultimately the list will contain the true model. The assumptions (A1.3) and the second part of (A1.4), used in Zheng and Loh (1997), imply in particular that with probability tending to one $(\mathbf{X}'\mathbf{X})^{-1}$ exists and therefore $\hat{\boldsymbol{\beta}}$ is unique. Similar conditions are used by Mammen (1993) to study the asymptotic behaviour of bootstrap estimators of contrasts in linear models of increasing dimension. We will consider in detail the case when M_m^n and M_M^n are optimised over all subsets of f i.e. $\mathcal{M} = 2^f$ and comment on the situation when the nested list of models is considered: $\mathcal{M}_{nested} = \{\{1, 2, \ldots, i\}\}_{i=1,\ldots,M_n}$. The first result concerns consistency of the minimal p-value criterion.

Theorem 1 *Let $\mathcal{M} = 2^f$. Then under conditions (A0), (A1.1'), (A1.2), (A1.3), (A1.4), (A1.5)* $P(M_m^n = t) \to 1,$ *as* $n \to \infty$.

As it follows from the proof an Lemma 4 condition (A1.1') may be weakened in Theorem 1 to $(a_n + \log n - M_n)/\sqrt{M_n} \to \infty$. We state now analogous result for MPVC criterion.

Theorem 2 *Let $\mathcal{M} = 2^f$. Then under conditions of Theorem 1 with (A1.1') replaced by (A1.1")* $P(M_M^n = t) \to 1,$ *as* $n \to \infty$.

In order to compare assumptions of the above results note that when M_n grows more slowly than $\log(n)$ we can take $a_n = 0$ in the case of criterion M_m^n. However, in the case of M_M^n the assumption (A1.1") is obviously not satisfied for $a_n = 0$.

It follows from the proof that the condition (A1.1") may be weakened in Theorem 1 to $(a_n - M_n)/\sqrt{M_n} \to \infty$.

Proofs of Theorems 1 and 2 are given in Section 4.

Consider now the case when the criteria are optimised over nested list of models $\mathcal{M}_{nested} = \{\{1, 2, \ldots, i\}\}_{i=1,\ldots,M_n}$ and define $i_{\max} = \max_{i\in t} i$ as the largest index of nonzero coefficient in the true model. In this case our goal is not to identify consistently the minimal true model t but rather $i_{\max}$, which is equivalent to consistent selection of a set $t_{\max} = \{1, \ldots, i_{\max}\}$. It turns out that this property holds under weaker conditions than in Theorem 1 and 2. Namely, the conditions (A1.3) and (A1.4) can be omitted. In this case the condition (A0) will be slightly modified. Let $\mathbf{x}_{lt_{\max}} = (x_{l,1}^{(n)}, \ldots, x_{l,i_{\max}}^{(n)})'$ be a vector of variables which pertain to the model $\{1, \ldots, i_{\max}\}$. Instead of (A0) we assume (B0): $\mathbf{E}(\mathbf{x}_{lt_{\max}}\mathbf{x}'_{lt_{\max}})$ is positive definite matrix. Then under conditions (B0), (A1.1'), (A1.2)and (A1.5) $P(M_m^n = t_{\max}) \to 1$ and analogous result holds for M_m^n provided (A1.1') is replaced by (A1.1"). This is proved along the lines of the proofs of Theorems 1 and 2.

In order to lessen computational burden of all subset search we propose two-step model selection with the first step consisting in initial ordering of variables according to p-values of coefficient of partial determination (3). This method is analogous to the procedure proposed in Zheng and Loh (1997) in which variables are ordered according to absolute values of t-statistics corresponding to respective attributes. Then in the second step an arbitrary criterion Crit is optimised over nested family of models. Specifically, the greedy procedure consists of the following steps. Let

$$PV_i = p(R^n_{(f-\{i\})f}|M_n, M_n - 1), \quad i = 1, \ldots, M_n \tag{5}$$

be the p-value of statistic $R^n_{(f-\{i\})f}$ for testing H_0 : model $f - \{i\}$ against H_1 : model f. Then

(Step 1) Order the p-values in nondecreasing order $PV_{i_1} \leq PV_{i_2} \leq \ldots \leq PV_{i_{M_n}}$.

(Step 2) Consider the nested family $\{\{i_1, i_2, \ldots, i_k\}\}_{k=1,\ldots,M_n}$ and optimise criterion Crit over this family.

It can be shown that under (A1.2)-(A1.4)

$$\lim_{n\to\infty} P(\max_{i\in t} PV_i < \min_{i\notin t} PV_i) = 1.$$

The proof of the above assertion is a simple consequence of Theorem 2 in Zheng and Loh (1997). This, together with Theorems 1 and 1 for the case of the nested list of models, when minimal or maximal p-value criterion is considered as Crit, leads to the following corollary.

Corollary 1 *Under conditions of Theorems 1 and 2 respectively the greedy versions of mPVC and MPVC procedures are consistent.*

Observe that since parameters of beta distribution used to calculate p-values in (5) do not change with i, the ordering in the first step is equivalent to ordering wrt values of $R^n_{(f-\{i\})f}$, or to the ordering wrt to absolute values of t-statistics when the full model is fitted.

3.2 Deterministic-design regression

In this section we will briefly discuss the case when the design matrix $\mathbf{X}$ is nonrandom. We allow that the values of attributes $\mathbf{x}^{(n)}_{l,1}, \dots, \mathbf{x}^{(n)}_{l,M_n}$ of l^{th} observation may depend on n. Recall that $\mathbf{x}_{lt} = \mathbf{x}^{(n)}_{lt}$ is a vector of variables which pertain to the minimal true model t. In the case of all subset search we replace condition (A0) by the following assumption

(C0) $n^{-1}\sum_{i=1}^{n} \mathbf{x}_{lt}\mathbf{x}'_{lt} \to \bar{\mathbf{W}}$, as $n \to \infty$, where $\bar{\mathbf{W}}$ is a positive definite matrix.

In the case of random covariates the above convergence in probability follows from The Law of Large Numbers. We also replace conditions (A1.3) and (A1.4) by the following assumption

(C1) The minimum eigenvalue $\tilde{\kappa}_n$ of $n^{-1}\mathbf{X}'\mathbf{X}$ is bounded away from zero, i.e. $\tilde{\kappa}_n > \tilde{\kappa} > 0$ for some $\tilde{\kappa} > 0$ and $n \in \mathbf{N}$.

Recall that $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_{M_n})'$ is the least squares estimator based on the full model f. Let $T_i = \hat{\sigma}^{-1}[(\mathbf{X}'\mathbf{X})^{-1}_{i,i}]^{-1/2}$ be the corresponding t-statistic. It can be easily shown that $\hat{\sigma}T_i = \beta_i[(\mathbf{X}'\mathbf{X})^{-1}_{i,i}]^{-1/2} + o_P(1)$, for $i \in t$. Thus by assumption (C1) $P(\hat{\sigma}T_i > Cn^{-1/2}) \to 1$ as $n \to \infty$, for some $C > 0$. This implies the conclusion of Lemma 5 in Section 4, namely that for $i \in t$ with probability tending to one $RSS(f - \{i\})/RSS(f)$ is bounded away from 0. As (A1.3) and (A1.4) are used in the random-design case only to prove Lemma 5 it follows that the analogous results to Theorem 1 and Theorem 2 hold for the deterministic-design case.

Corollary 2 *Under conditions (C0), (A1.1'), (A1.2), (C1), (A1.5)*
$P(M^n_m = t) \to 1$, *as* $n \to \infty$.

Corollary 3 *Under conditions of Corollary 2 with (A1.1') replaced by (A1.1")*
$P(M^n_M = t) \to 1$, *as* $n \to \infty$.

Consider the case of nested family search. Recall that $\mathbf{x}_{lt_{\max}}$ is a vector of variables which pertain to the model $\{1, \dots, i_{\max}\}$. If condition (B0) if replaced by the following assumption

(D0) $n^{-1}\sum_{i=1}^{n} \mathbf{x}_{lt_{\max}}\mathbf{x}'_{lt_{\max}} \to \tilde{\mathbf{W}}$, as $n \to \infty$, where $\tilde{\mathbf{W}}$ is a positive definite matrix.

then results discussed at the end of Section 3.1 hold for deterministic design.

4 Proofs

We first state auxiliary lemmas which will be used in the proof of Theorem 1. The first one proved in Pokarowski and Mielniczuk (2010) gives an approximation of tail probability function of beta distribution. Let $B_{a,b}$ be a random variable having beta distribution with shape parameters a and b and $B(x,y)$ denote beta function. Define an auxiliary function

$$L(a,b,x) = \frac{(a-1)(1-x)}{1-a+(a+b)x},$$

for $a,b,x \in \mathbf{R}$ such that $x \neq (a-1)/(a+b)$.

Lemma 1 *Assume $x > \frac{a-1}{a+b}$. Then for $a \geq 1$*

$$\frac{(1-x)^b x^{a-1}}{B(a,b)b} \leq P[B_{a,b} > x] \leq \frac{(1-x)^b x^{a-1}}{B(a,b)b}(1+L(a,b,x)) \tag{6}$$

and for $a < 1$

$$\frac{(1-x)^b x^{a-1}}{B(a,b)b}(1+L(a,b,x)) \leq P[B_{a,b} > x] \leq \frac{(1-x)^b x^{a-1}}{B(a,b)b}. \tag{7}$$

The following Lemma states simple but useful inequalities for gamma function.

Lemma 2 *Let $a = p/2$ and $b = (n-p)/2$, for some $p, n \in \mathbf{N}$. Then*

$$\Gamma(b)b^a \leq \Gamma(a+b) \leq \frac{2}{\sqrt{\pi}}\Gamma(b)(a+b)^a.$$

The above Lemma implies an inequality for beta function $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a,b)$

$$\frac{b^{a-1}}{\Gamma(a)} \leq \frac{1}{bB(a,b)} \leq \frac{2}{\sqrt{\pi}}\frac{(a+b)^a}{b\Gamma(a)}, \tag{8}$$

for $a = p/2$, $b = (n-p)/2$ and $p, n \in \mathbf{N}$.

Remark 1 *Lemma 2 easily implies inequality $\Gamma(p/2) \leq (\lceil p/2 \rceil - 1)! \leq p^{p/2}$ for $p > 1$, which will be frequently used throughout.*

The following Lemma states that for a proper submodel of the true model t variance estimator is asymptotically biased. $j \subset k$ denotes a proper inclusion of j in k.

Lemma 3 *(i) For $j \supseteq t$, $j \in \mathcal{M}$ $\frac{RSS(j)}{n} \xrightarrow{P} \sigma^2$ as $n \to \infty$. Moreover, for $j \subset t$, $j \in \mathcal{M}$ if (A0) is satisfied then $\frac{RSS(j)}{n} \xrightarrow{P} \sigma^2 + \lambda_j$ as $n \to \infty$, where $\lambda_j > 0$.*
(ii) Let $j \subset t_{\max}$, $j \in \mathcal{M}_{nested}$ and assume (B0). Then $\frac{RSS(j)}{n} \xrightarrow{P} \sigma^2 + \lambda_j$ as $n \to \infty$, where $\lambda_j > 0$.

Lemma 4 *Let R_n be a sequence of real numbers such that $(R_n - M_n)/\sqrt{M_n} \to \infty$ as $n \to \infty$. Assume also that $M_n/n \to 0$ and matrix $\mathbf{X}'\mathbf{X}$ is invertible with probability tending to 1. Then*

$$P\left\{n \log\left[\frac{RSS(t)}{RSS(f)}\right] > R_n\right\} \to 0$$

as $n \to \infty$.

Remark 2 *Observe that as $(R_n - M_n)/\sqrt{M_n} = \sqrt{M_n}(R_n/M_n - 1)$, the imposed condition on R_n is implied by $R_n/M_n \to \infty$. Thus in particular Lemma 4 implies that*

$$\frac{RSS(t)}{RSS(f)} = O_P\left[\exp\left(\frac{R_n}{n}\right)\right],$$

for any R_n such that $R_n/M_n \to \infty$. Observe moreover that Lemma 4 holds true also in the case $M_n = M$ when the condition on R_n reduces to $R_n \to \infty$ only and thus $RSS(t)/RSS(f) = O_P(exp(n^{-1}))$. This can be seen directly from Lemma 3 and the fact that $RSS(t) - RSS(f) \sim \chi^2_{M-p_t}$ as it follows from them that $R^n_{tf} = \mathcal{O}_P(n^{-1})$ and thus $n\log(RSS(t)/RSS(f)) = \mathcal{O}_P(1)$.

Lemma 5 *Assume conditions (A1.3) and (A1.4). Then there exists $a > 0$ such that*

$$P\left\{\min_{i\in t} \log\left[\frac{RSS(f-\{i\})}{RSS(f)}\right] > a\right\} \to 1$$

as $n \to \infty$.

Thus Lemma 5 implies that with probability tending to 1 $RSS(f-\{i\})/RSS(f)$ for $i \in t$ is bounded away from 0.

4.1 Proof of Theorem 1

We will consider separately two cases: the first when the true model t contains nontrivial regressors ($p_t \geq 1$) and the second, when it equals the null model.

Case 1 ($p_t \geq 1$). We will treat the case $p_t \geq 2$ in detail, the case $p_t = 1$ is similar but simpler and relies on (7) instead of (6) to treat $p(R^n_{0t}|p_t, 0)$.

(i) Let j be such that $j \supset t$ i.e. t is a proper subset of j. We will prove that $P[e^{p_t a_n} p(R^n_{0t}|p_t, 0) > \inf_{j\supset t} e^{p_j a_n} p(R^n_{0j}|p_j, 0)] \to 0$ as $n \to \infty$. Using (8) with $a = p_t/2$ and $b = (n-p_t)/2$ we obtain the following inequalities for sufficiently large n

$$\frac{1}{B(\frac{p_t}{2}, \frac{n-p_t}{2})\left(\frac{n-p_t}{2}\right)} \leq \frac{2\left(\frac{n}{2}\right)^{\frac{p_t}{2}}}{\sqrt{\pi}\left(\frac{n-p_t}{2}\right)\Gamma\left(\frac{p_t}{2}\right)} \leq \frac{2\left(\frac{n}{2}\right)^{\frac{p_t}{2}}}{\sqrt{\pi}\left(\frac{n}{4}\right)\Gamma\left(\frac{p_t}{2}\right)} = \frac{4\left(\frac{n}{2}\right)^{\frac{p_t}{2}-1}}{\sqrt{\pi}\Gamma\left(\frac{p_t}{2}\right)}. \tag{9}$$

Moreover for $j \supset t$ and sufficiently large n

$$\frac{1}{B(\frac{p_j}{2},\frac{n-p_j}{2})\left(\frac{n-p_j}{2}\right)} \geq \frac{\left(\frac{n-p_j}{2}\right)^{\frac{p_j}{2}-1}}{\Gamma\left(\frac{p_j}{2}\right)} \geq \frac{\left(\frac{n-M_n}{2}\right)^{\frac{p_j}{2}-1}}{M_n^{\frac{p_j}{2}}} \geq \frac{\left(\frac{n-M_n}{2}\right)^{\frac{p_t+1}{2}-1}}{M_n^{\frac{p_t+1}{2}}} \geq \frac{\left(\frac{n}{2}\right)^{\frac{p_t+1}{2}-1}\left(\frac{1}{2}\right)^{\frac{p_t+1}{2}-1}}{M_n^{\frac{p_t+1}{2}}}. \tag{10}$$

Note that

$$P\left(\inf_{j\supset t} R_{0j}^n \geq \sup_{j\supset t}\frac{\frac{p_j}{2}-1}{\frac{n}{2}}\right) \leq P\left(R_{0t}^n \geq (M_n-2)/n\right) \to 1,$$

which follows from Lemma 3 and the fact that $M_n/n \to 0$. Thus the assumption of Lemma 1 is satisfied for $x = R_{0j}^n$, $a = \frac{p_j}{2}$, $b = \frac{n-p_j}{2}$ and all $j \supset t$. Using (6) we have

$$\begin{aligned}
&P[e^{p_t a_n} p(R_{0t}^n|p_t,0) > \inf_{j\supset t} e^{p_t a_n} p(R_{0j}^n|p_j,0)] \leq \\
&P\left\{\frac{(1-R_{0t}^n)^{\frac{n-p_t}{2}}(R_{0t}^n)^{\frac{p_t}{2}-1}[1+L\left(\frac{p_t}{2},\frac{n-p_t}{2},R_{0t}^n\right)]e^{p_t a_n}}{B(\frac{p_t}{2},\frac{n-p_t}{2})\left(\frac{n-p_t}{2}\right)} > \inf_{j\supset t}\frac{(1-R_{0j}^n)^{\frac{n-p_j}{2}}(R_{0j}^n)^{\frac{p_j}{2}-1}e^{(p_t+1)a_n}}{B(\frac{p_j}{2},\frac{n-p_j}{2})\left(\frac{n-p_j}{2}\right)}\right\} \leq \\
&P\left\{\frac{(1-R_{0t}^n)^{\frac{n-p_t}{2}}[1+L\left(\frac{p_t}{2},\frac{n-p_t}{2},R_{0t}^n\right)]e^{p_t a_n}}{B(\frac{p_t}{2},\frac{n-p_t}{2})\left(\frac{n-p_t}{2}\right)} > \inf_{j\supset t}\frac{(1-R_{0f}^n)^{\frac{n-p_t}{2}}(R_{0t}^n)^{\frac{M_n}{2}-1}e^{(p_t+1)a_n}}{B(\frac{p_j}{2},\frac{n-p_j}{2})\left(\frac{n-p_j}{2}\right)}\right\}.
\end{aligned} \tag{11}$$

Taking logarithms and using inequalities (9), (10) we obtain

$$P[\log p(R_{0t}^n|p_t,0) + p_t a_n > \inf_{j\supset t}\log p(R_{0j}^n|p_j,0) + (p_t+1)a_n] \leq P\left\{\left[\frac{n-p_t}{2}\right]\log\left[\frac{RSS(t)}{RSS(f)}\right] > \tilde{W}_n\right\},$$

where

$$\begin{aligned}
\tilde{W}_n &= a_n + \frac{1}{2}\log\left(\frac{n}{2}\right) - \log[1+L\left(\frac{p_t}{2},\frac{n-p_t}{2},R_{0t}^n\right)] + \left(\frac{M_n}{2}-1\right)\log(R_{0t}^n) + \\
&\left(\frac{p_t+1}{2}-1\right)\log\left(\frac{1}{2}\right) - \left(\frac{p_t+1}{2}\right)\log(M_n) - \log\left(\frac{4}{\sqrt{\pi}}\right) + \log\Gamma\left(\frac{p_t}{2}\right).
\end{aligned}$$

Assumption $M_n/(a_n+\log(n)) \to 0$, Lemma 3 and the fact that $R_{0,t} \xrightarrow{P} \sigma^2 > 0$ imply that there exists a sequence W_n of real numbers such that $P(\tilde{W}_n > W_n) \to 1$ and $W_n/M_n \to \infty$. Now the required convergence follows from

$$P\left\{\left[\frac{n-p_t}{2}\right]\log\left[\frac{RSS(t)}{RSS(f)}\right] > W_n\right\} \to 0$$

which in its turn is implied by Lemma 4.

(ii) Consider now the case $j \not\supseteq t$ and let $i = i(j) \in \mathbf{N}$ be such that $i \in t \cap j^c$. We will prove that $P[e^{p_t a_n} p(R_{0t}^n|p_t,0) > \inf_{j\not\supseteq t} e^{p_j a_n} p(R_{0j}^n|p_j,0)] \to 0$ as $n \to \infty$. Define $M(n,i) = \max\{R_{0(f-\{i\})}^n, \frac{2M_n}{(n-M_n)}\}$,

for $i \in t$. Assume first that $p_j \geq 2$. Using (6) and (8) we have

$$e^{p_j a_n} p(R_{0j}^n | p_j, 0) \geq e^{2a_n} p(M(n,i)|p_j, 0) \geq \frac{e^{2a_n}[1 - M(n,i)]^{\frac{n-p_j}{2}} M(n,i)^{\frac{p_j}{2}-1}}{B\left(\frac{p_j}{2}, \frac{n-p_j}{2}\right)\left(\frac{n-p_j}{2}\right)} \geq$$
$$\frac{e^{2a_n}[1 - M(n,i)]^{\frac{n}{2}} \left(\frac{2M_n}{n-M_n}\right)^{\frac{p_j}{2}-1}}{B\left(\frac{p_j}{2}, \frac{n-p_j}{2}\right)\left(\frac{n-p_j}{2}\right)} \geq \frac{e^{2a_n}[1 - M(n,i)]^{\frac{n}{2}} \left(\frac{2M_n}{n-M_n}\right)^{\frac{p_j}{2}-1} \left(\frac{n-M_n}{2}\right)^{\frac{p_j}{2}-1}}{\Gamma\left(\frac{p_j}{2}\right)} \geq$$
$$\frac{e^{2a_n}[1 - M(n,i)]^{\frac{n}{2}} M_n^{\frac{p_j}{2}-1}}{M_n^{\frac{p_j}{2}}} = e^{2a_n}[1 - M(n,i)]^{\frac{n}{2}} M_n^{-1}. \tag{12}$$

From (6) and (9)

$$e^{p_t a_n} p(R_{0t}^n | p_t, 0) \leq \frac{e^{p_t a_n}(1 - R_{0t}^n)^{\frac{n-p_t}{2}} 4\left(\frac{n}{2}\right)^{\frac{p_t}{2}-1}\left[1 + L\left(\frac{p_t}{2}, \frac{n-p_t}{2}, R_{0t}^n\right)\right]}{\sqrt{\pi}\Gamma\left(\frac{p_t}{2}\right)} \tag{13}$$

Using (12) and (13) we have for $p_t \geq 2$ and $p_j \geq 2$

$$P[e^{p_t a_n} \log p(R_{0t}^n | p_t, 0) > \inf_{j \not\supseteq t} e^{p_j a_n} \log p(R_{0j}^n | p_j, 0)] \leq P\left\{\inf_{i \in t} \frac{n}{2} \log\left[\frac{(1 - M(n,i))RSS(0)}{RSS(t)}\right] < \tilde{S}_n\right\},$$

where

$$\tilde{S}_n = a_n(p_t - 2) + \left(\frac{p_t}{2} - 1\right)\log\left(\frac{n}{2}\right) - \frac{p_t}{2}\log\left(\frac{RSS(t)}{RSS(0)}\right) + \log\left(\frac{4}{\sqrt{\pi}}\right) +$$
$$\log\left[1 + L\left(\frac{p_t}{2}, \frac{n-p_t}{2}, R_{0t}^n\right)\right] + \log\Gamma^{-1}\left(\frac{p_t}{2}\right) + \log(M_n).$$

In view of definition of $M(n,i)$ the last probability can be bounded from above by

$$P\left\{\inf_{i \in t} \frac{n}{2} \log\left[\frac{RSS(f - \{i\})}{RSS(t)}\right] < \tilde{S}_n\right\} + P\left\{\frac{n}{2}\log\left[\frac{(1 - \frac{2M_n}{n-M_n})RSS(0)}{RSS(t)}\right] < \tilde{S}_n\right\}.$$

The second probability above converges to zero in view of Lemma 3. Consider the first probability. Since the number of elements of t is finite it suffices show that $P\left\{\frac{n}{2}\log\left[\frac{RSS(f-\{i\})}{RSS(t)}\right] < \tilde{S}_n\right\} \to 0$ for any $i \in t$. Namely, it is bounded from above by

$$P\left\{\frac{n}{2}\log\left[\frac{RSS(f - \{i\})}{RSS(f)}\right] + \frac{n}{2}\log\left[\frac{RSS(f)}{RSS(t)}\right] < \tilde{S}_n\right\} \leq$$
$$P\left\{\frac{n}{2}\log\left[\frac{RSS(f - \{i\})}{RSS(f)}\right] < 2\tilde{S}_n\right\} + P\left\{\frac{n}{2}\log\left[\frac{RSS(f)}{RSS(t)}\right] < -\tilde{S}_n\right\} \leq$$
$$P\left\{n\log\left[\frac{RSS(f - \{i\})}{RSS(f)}\right] < \tilde{S}_n\right\} + P\left\{\frac{n}{2}\log\left[\frac{RSS(t)}{RSS(f)}\right] \geq \tilde{S}_n\right\}. \tag{14}$$

From assumptions (A1.5) and (A1.1') $\tilde{S}_n/n \xrightarrow{P} 0$ and $\tilde{S}_n/M_n \xrightarrow{P} \infty$, respectively. Thus the convergence to zero of the above two probabilities in (14) follows from Lemma 5 and 4, respectively. The case $p_j = 1$ is treated analogously.

Consider now the case $p_j = 0$. From (13) we have

$$\begin{aligned}
&P[\log p(R^n_{0t}|p_t, 0) + p_t a_n > \log p(R^n_{00}|0, 0)] = P[\log p(R^n_{0t}|p_t, 0) > a_n - \frac{1}{2}\log(n) - p_t a_n] \leq \\
&P\left\{\left(\frac{n-p_t}{2}\right)\log\left[\frac{RSS(0)}{RSS(t)}\right] < G_n\right\}, \qquad (15)
\end{aligned}$$

where

$$\begin{aligned}
G_n = &(p_t - 1)a_n + \frac{1}{2}\log(n) + \left(\frac{p_t}{2} - 1\right)\log\left(\frac{n}{2}\right) + \log\left(\frac{4}{\sqrt{\pi}}\right) + \log\Gamma^{-1}\left(\frac{p_t}{2}\right) + \\
&\log\left[1 + L\left(\frac{p_t}{2}, \frac{n-p_t}{2}, R^n_{0t}\right)\right].
\end{aligned}$$

The convergence to zero of the probability in (15) follows from Lemma 3 amd assumption (A1.5).

Case 2 ($p_t = 0$) i.e. the true model is null model. We treat in detail the case $p_j \geq 2$. Define $\bar{M}(n) = \max\{R^n_{0f}, \frac{2M_n}{n-M_n}\}$. Note that the assumption of Lemma 1 is satisfied for $x = \bar{M}(n)$, $a = \frac{p_j}{2}$, and $b = \frac{n-p_j}{2}$. Using (6) and (8) we have

$$\begin{aligned}
&e^{p_j a_n} p(R^n_{0j}|p_j, 0) \geq e^{2a_n} p(\bar{M}(n)|p_j, 0) \geq \frac{e^{2a_n}[1-\bar{M}(n)]^{\frac{n-p_j}{2}}\bar{M}(n)^{\frac{p_j}{2}-1}}{B\left(\frac{p_j}{2}, \frac{n-p_j}{2}\right)\left(\frac{n-p_j}{2}\right)} \geq \\
&\frac{e^{2a_n}[1-\bar{M}(n)]^{\frac{n-p_t}{2}}\left(\frac{2M_n}{n-M_n}\right)^{\frac{p_j}{2}-1}}{B\left(\frac{p_j}{2}, \frac{n-p_j}{2}\right)\left(\frac{n-p_j}{2}\right)} \geq \frac{e^{2a_n}[1-\bar{M}(n)]^{\frac{n-p_t}{2}}\left(\frac{2M_n}{n-M_n}\right)^{\frac{p_j}{2}-1}\left(\frac{n-M_n}{2}\right)^{\frac{p_j}{2}-1}}{\Gamma\left(\frac{p_j}{2}\right)} \geq \\
&\frac{e^{2a_n}[1-\bar{M}(n)]^{\frac{n-p_t}{2}} M_n^{\frac{p_j}{2}-1}}{M_n^{\frac{p_j}{2}}} = e^{2a_n}[1-\bar{M}(n)]^{\frac{n-p_t}{2}} M_n^{-1}. \qquad (16)
\end{aligned}$$

Using (16) we obtain the following inequality

$$\begin{aligned}
&P[\log p(R^n_{00}|0, 0) > \inf_{j:p_j \geq 2}\log p(R^n_{0j}|p_j, 0) + 2a_n] \leq P[a_n - \frac{1}{2}\log(n) > \inf_{j:p_j \geq 2}\log p(R^n_{0j}|p_j, 0) + 2a_n] \leq \\
&P\left\{-\left(\frac{n-p_t}{2}\right)\log[1-\bar{M}(n)] > a_n + \frac{1}{2}\log(n) - \log(M_n)\right\} \leq \\
&P\left\{\left(\frac{n-p_t}{2}\right)\log\left[\frac{RSS(0)}{RSS(f)}\right] > a_n + \frac{1}{2}\log(n) - \log(M_n)\right\} + \\
&I\left[-\left(\frac{n-p_t}{2}\right)\log\left(1 - \frac{2M_n}{n-M_n}\right) > a_n + \frac{1}{2}\log(n) - \log(M_n)\right]. \qquad (17)
\end{aligned}$$

From Lemma 4 and the assumption $M_n/(a_n + \log(n)) \to 0$ the first probability in (17) converges to zero. The same assumption implies that the second term is ultimately 0. This completes the proof.

4.2 Proof of Theorem 2

The proof is similar to that of Theorem 1 and splits into two cases: $M_n - p_t \geq 1$ (corresponding to the case $p_t \geq 1$ in the previous proof) and $M_n = p_t$ (corresponding to the former case $p_t = 0$). We give the sketch of the proof only.

Case 1 ($M_n - p_t \geq 1$). We discuss the situation when $M_n - p_t \geq 2$, the remaining case relies on (7) instead of (6). Define $\tilde{M}(n,t) = \max\{R^n_{tf}, \frac{2M_n}{n-M_n}\}$. Note that the assumption of Lemma 1 is satisfied for $x = \tilde{M}(n,t)$, $a = \frac{M_n - p_t}{2}$, and $b = \frac{n-M_n}{2}$. In this case condition $a \geq 1$ is also satisfied. Analogously to the proof of (16) we obtain

$$p(R^n_{tf}|M_n, p_t) \geq p(\tilde{M}(n,t)|M_n, p_t) \geq [1 - \tilde{M}(n,t)]^{\frac{n-M_n}{2}} M_n^{-1}. \tag{18}$$

(i) Let j be such that $j \supset t$ i.e. t is a proper subset of j. We will prove that $P[e^{-p_t a_n} p(R^n_{tf}|M_n, p_t) < \sup_{j \supset t} e^{-p_j a_n} p(R^n_{jf}|M_n, p_j)] \to 0$ as $n \to \infty$. For $j \supset t$ we have $e^{-p_j a_n} p(R^n_{jf}|M_n, p_j) \leq \exp[-(p_t+1)a_n]$. This inequality also applies to $j = f$. Thus using (18) we obtain the following inequalities

$$\begin{aligned}
&P[e^{-p_t a_n} p(R^n_{tf}|M_n, p_t) < \sup_{j \supset t} e^{-p_j a_n} p(R^n_{jf}|M_n, p_j)] \leq \\
&P\left\{\left(\frac{n-M_n}{2}\right) \log[1 - \tilde{M}(n,t)] - \log(M_n) - p_t a_n < -(p_t+1)a_n\right\} \leq \\
&P\left\{\left(\frac{n-M_n}{2}\right) \log\left[\frac{RSS(t)}{RSS(f)}\right] > a_n - \log(M_n)\right\} + \\
&I\left\{-\left(\frac{n-M_n}{2}\right) \log\left[1 - \frac{2M_n}{n-M_n}\right] > a_n - \log(M_n)\right\}.
\end{aligned}$$

The above bound converges to zero in view of the assumption $M_n/a_n \to 0$ and Lemma 4.

(ii) Consider now the case $j \not\supseteq t$ and assume that $p_j \leq M_n - 2$ (this corresponds to $p_j \geq 2$ in the previous proof). Let index $i = i(j)$ be such that $i \in t \cap j^c$. It follows from Lemma 5 that the assumption of Lemma 1 is satisfied for $x = R_{(f-\{i\})f}$, $a = \frac{M_n - p_j}{2}$, and $b = \frac{n-M_n}{2}$. Moreover the same reasoning yields for all $j \not\supseteq t$ $L\left(\frac{M_n-p_j}{2}, \frac{n-M_n}{2}, R_{f-\{i\}f}\right) \leq M_n$ wih probability tending to 1. Using (6) we have the following inequalities

$$\begin{aligned}
&e^{-p_j a_n} p(R^n_{jf}|M_n, p_j) \leq p(R_{(f-\{i\})f}|M_n, p_j) \leq \\
&\frac{[1 - R_{(f-\{i\})f}]^{\frac{n-M_n}{2}} [R_{(f-\{i\})f}]^{\frac{M_n-p_j}{2}-1}}{B\left(\frac{M_n-p_j}{2}, \frac{n-M_n}{2}\right)\left(\frac{n-M_n}{2}\right)} \left[1 + L\left(\frac{M_n-p_j}{2}, \frac{n-M_n}{2}, R_{(f-\{i\})f}\right)\right] \leq \\
&[1 - R_{(f-\{i\})f}]^{\frac{n-M_n}{2}} \frac{2n^{\frac{M_n}{2}}}{\sqrt{\pi}\Gamma\left(\frac{M_n}{2}\right)} [1 + M_n].
\end{aligned} \tag{19}$$

Thus

$$P[e^{-p_t a_n} p(R^n_{tf}|M_n, p_t) < \sup_{j \not\supseteq t} e^{-p_j a_n} p(R^n_{jf}|M_n, p_j)] \leq$$

$$P\left\{\sup_{i\in t}\left(\frac{n-M_n}{2}\right)\log\left[\frac{\left(1-\tilde{M}(n,t)\right)RSS(f-\{i\})}{RSS(f)}\right]<K_n\right\},$$

where

$$K_n=p_ta_n+\log\left(\frac{2}{\sqrt{\pi}}\right)+\frac{M_n}{2}\log(n)-\log\Gamma\left(\frac{M_n}{2}\right)+\log(1+M_n)+\log(M_n).$$

Similarly to the proof of (14) we obtain that the RHS tends to 0.
The case $p_j > M_n - 2$ is simpler and uses (7) instead of (6).
Case 2 ($M_n = p_t$). Thus $e^{-p_ta_n}p(R^n_{tf}|M_n,p_t) = e^{-M_na_n}$. Assume $p_j \le M_n - 2$ and let $i = i(j)$ be such that $i \in j^c \cap t$. Then using $L\left(\frac{M_n-p_j}{2},\frac{n-M_n}{2},R_{jf}\right) \le M_n$ and (6) (cf (19) it is easy to establish that

$$e^{-p_ja_n}p(R^n_{jf}|M_n,p_j)\le p(R_{(f-\{i\})f}|M_n,p_j)\le[1-R_{(f-\{i\})f}]^{\frac{n-M_n}{2}}\frac{2n^{\frac{M_n}{2}}}{\sqrt{\pi}\Gamma\left(\frac{M_n}{2}\right)}[1+M_n].$$

Then it follows that

$$\begin{aligned}&P[e^{-p_ta_n}p(R^n_{tf}|M_n,p_t)<\sup_{j\neq t}e^{-p_ja_n}p(R^n_{jf}|M_n,p_j)]\le\\&P\left\{\sup_{i\in t}\left(\frac{n-M_n}{2}\right)\log\left[\frac{RSS(t-\{i\})}{RSS(t)}\right]<\tilde{K}_n\right\},\end{aligned}$$

where

$$\tilde{K}_n=M_na_n+\log(2/\sqrt{\pi})+\frac{M_n}{2}\log(n)-\log\Gamma(M_n/2)+\log(1+M_n)+\log(M_n).$$

The convergence to zero of the above probability follows from Lemma 3 and the assumption $a_n/n \to 0$. The case $p_j > M_n - 2$ is analogous.

5 Numerical experiments

In this section we study the finite-sample performance of the model selection procedures. We consider criteria defined in Section 2: minimal p-value criterion M^n_m with $a_n = 0$ which will be called simply in this section mPVC and two scaled p-value criteria with scalings which were empirically chosen, namely minimal p-value criterion with $a_n = \log(n)/2$ and maximal p-value criterion with the same a_n called mPVCcal and MPVCcal, respectively. As benchmarks we considered performance of classical criteria based on penalized log-likelihood which have the form

$$\text{argmax}_{j\in\mathcal{M}}\{2\log\mathbf{f}_{\hat{\boldsymbol{\beta}}_j,\hat{\sigma}^2_j}(\mathbf{Y}|\mathbf{X})-p_jC_n\}=\text{argmax}_{j\in\mathcal{M}}\{-n\log[RSS(j)/n]-p_jC_n\}$$

with penalties: $C_n = 2$ and $C_n = \log(n)$ which correspond to Akaike (AIC) and Bayesian (BIC) information criteria, respectively.

5.1 Simulation experiments

The simulation experiments were carried out with sample sizes $n = 75, 100, 200, 300, 500, 1000$ repeated $N = 500$ times. We consider the following lists of models

(M1) $t = \{10\}$, $\beta_1 = 0.2$, $M_n = 30$,

(M2) $t = \{1, 2, 5, 6\}$, $\boldsymbol{\beta} = (0.9, -0.8, -0.4, 0.2)'$, $M_n = 6$,

(M3) $t = \{2, 4, 5\}$, $\boldsymbol{\beta} = (1, 1, 1)'$, $M_n = 5$,

(M4) $t = \{2k + 7 : k = 3, \dots, 12\}$, $\boldsymbol{\beta} = (1, \dots, 1)'$, $M_n = 60$.

In all cases $\mathcal{M} = 2^{\{1,\dots,M_n\}}$. Models M1, M3 and M4 were also considered in Zheng and Loh (1997). Regressors $\mathbf{x}_l^n$ were generated from M_n-variate zero mean normal distribution with (i, j)th entry of the covariance matrix $\Sigma_X = (\sigma_{ij})_{ij}$ equal $\sigma_{ij} = 0.5^{|i-j|}$. The distribution of $(\varepsilon_1, \dots, \varepsilon_n)$ was multivariate standard normal. We considered greedy variants of the selection methods, described in Section 3. Table 1 presents estimated probabilities of correct ordering, e.g. the probabilities that the coordinates corresponding to nonzero coefficients are placed ahead the spurious ones. It is seen that for $n \geq 500$ for the models considered a correct ordering is recovered practically always. We assess the effectiveness of the selection rule in terms of the probability of true model selection $P(\hat{t} = t)$, where $\hat{t}$ is a model selected by the considered rule and mean squared error $\mathbf{E}(||\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}(\hat{t})||^2)$, where $\hat{\boldsymbol{\beta}}(\hat{t})$ is the post-model selection estimator of $\boldsymbol{\beta}$ i.e. ML estimator in the chosen model. In the experiments estimates of these measures calculated as the empirical means of respective quantities were considered. The influence of the sample size on the effectiveness of selected rules has been investigated. For models M1, M3 and M4 criterion MPVCcal and mPVCcal perform considerably better for all sample sizes considered than mPVC and commonly used BIC and AIC (see Figure 1 and 2). In contrast, in the case of model M2 criterion mPVC works better than others. In general, performance of mPVCcal is similar to that of MPVCcal. The results also indicate that model M1 with the only one significant variable placed at position 10 is the most difficult for selection among the models considered. This is due to the fact that in this case it is difficult to recover the correct ordering (see Table 1), especially for small sample sizes. Secondly the selection criteria seem to work worse when the number of nuisance covariates is large. For model M1 we also studied the influence of the value of the true parameter β_1. Figure 3 indicates that performance of both measures is much worse for small values of the parameter. The influence of the size of the list M_n on the effectiveness of selection rules has been also investigated. Figure 4 shows that for model M1 performance of the AIC, BIC and mPVC is influenced by the choice of the horizon M_n, however, the selection rules MPVCcal and mPVCcal are the least affected. We also investigated the influence of the strength of dependence structure of design matrix $\mathbf{X}$ on the behaviour of selection rules. We studied the cases when the dependence between the covariates is respectively stronger and weaker than in the case described above. Namely the covariances

$\Sigma_X(i,j) = 0.8^{|i-j|}$ and $\Sigma_X(i,j) = I\{i = j\}$ were considered. For the above cases we took also different marginal variances of regressors equal to 0.5 and 2. The error variance σ^2 was always set to one. The experiments show that the probability of true model selection is smaller (and respective prediction error larger) than for initial scenario when the dependence is stronger or the variance of covariates larger. However, it turns out that the ranking of methods with respect to both considered measures remains the same in all above cases. Experiments indicate also that for the considered selection criteria mean prediction error behaves approximately as a constant minus probability of a correct selection.

We also investigated the case of covariates $\mathbf{x}_l^n$ having different distributions. Namely, we considered the following regression scenario

$$\mathbf{Y} = \boldsymbol{\beta}'\mathbf{L}(\mathbf{U}) + \boldsymbol{\varepsilon},$$

where $\mathbf{L}(\cdot) = (L_1(\cdot), \ldots, L_{M_n}(\cdot))'$ is a vector consisting of the consecutive orthonormal Legendre polynomials on $[-1, 1]$ and $\mathbf{U}$ is random vector with continuous uniform distribution on $[-1, 1]$. We considered the following list of models

(L1) $t = \{1, 2, 4\}$, $\boldsymbol{\beta} = (1, 1, 1)'$

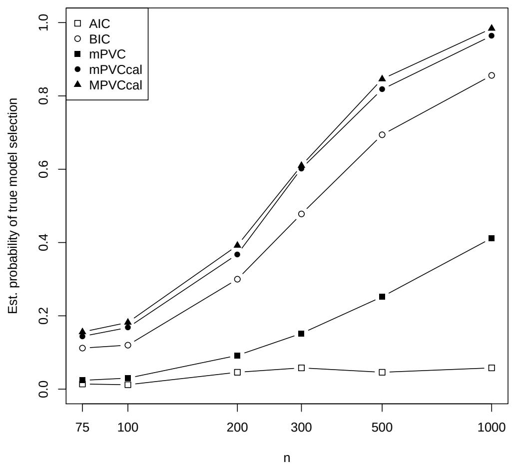
with horizons $M_n = 5, 10, \ldots, 25$. The influence of the size of the list M_n has been investigated. The sample size was set to $n = 300$. Figure 5 presents the results which are similar to that of the previous experiments indicating that mPVCcal and MPVCcal perform the best in this case, and the second best is BIC.

5.2 Real data example

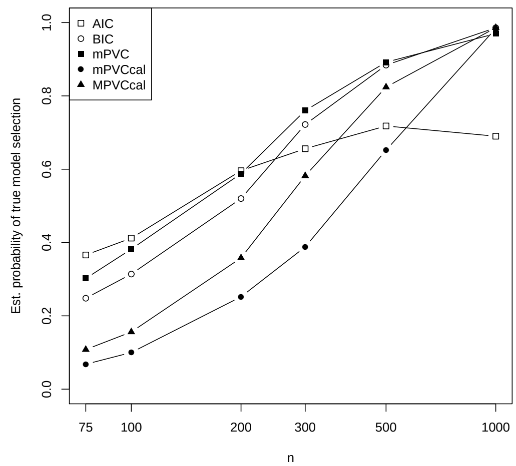
We consider `bodyfat` data set (Johnson (1996)) consisting of records of the percentage of fat in the body (dependent variable) together with 13 independent variables for $n = 252$ individuals. Two independent variables were selected having the smallest p-values when the full linear model was fitted. They were abdomen and wrist circumference and when used as predictors resulted in the fitted model with a vector of estimated coefficients $\hat{\boldsymbol{\beta}} = (0.7661, -2.8379)'$ and a variance of residuals $\hat{\sigma}^2 = 4.45$. A parametric bootstrap (see e.g. Davison and Hinkley (1997)) was employed to check how the considered selection criteria perform for this data set. Namely, the true model was the fitted linear model with the original two regressors, $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and the normal errors with the variance equal to $\hat{\sigma}^2$. Additional superfluous explanatory variables were created in pairs by drawing from the two-dimensional normal distribution with independent components, which mean and variance vector matched that of the original predictors. We considered $k = 8, 18, \ldots, 58$ additional variables what amounted to horizons $M_n = 10, 20, \ldots, 60$ when the true variables were accounted for. Thus M_n/n ranged from 0.03 to 0.23. 500 parametric bootstrap samples consisting of 252 observations each were created to mimic the original sample and the considered selection criteria were employed to choose subset of potential M_n variables. Figure 6 presents the results. The results are similar to that of simulation experiments indicating that mPVCcal and MPVCcal perform the best in this case, and the second best is BIC.

Table 1: Estimated probability of correct ordering based on $N = 500$ trials.

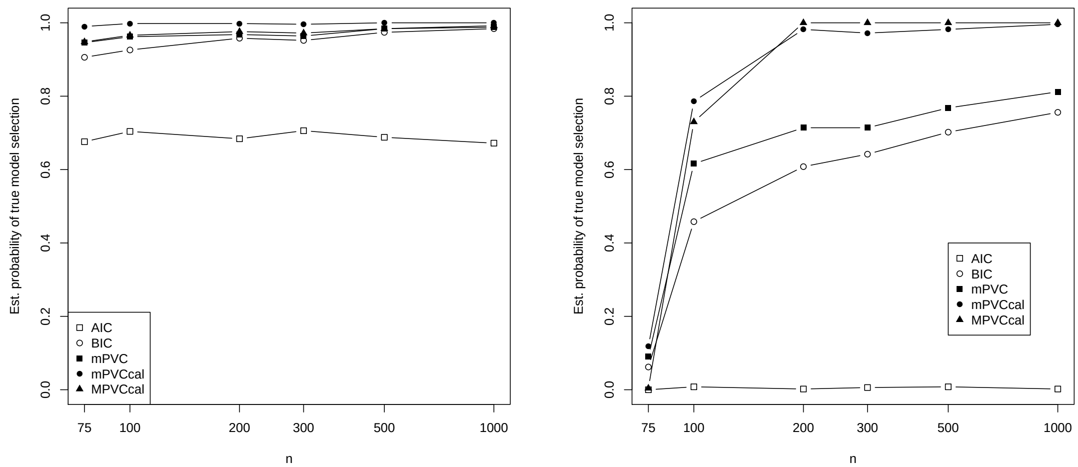
Model	$n = 75$	$n = 100$	$n = 200$	$n = 300$	$n = 500$	$n = 1000$
(M1)	0.16	0.18	0.39	0.61	0.85	0.98
(M2)	0.69	0.74	0.91	0.99	0.99	1
(M3)	0.99	1	1	1	1	1
(M4)	0.99	1	1	1	1	1
Est. max. standard error ≤ 0.01						



(a)

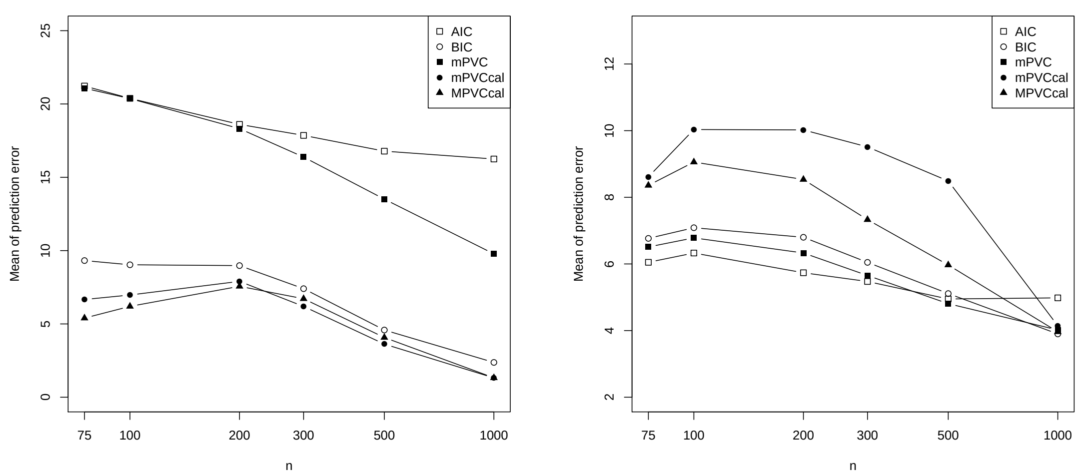


(b)

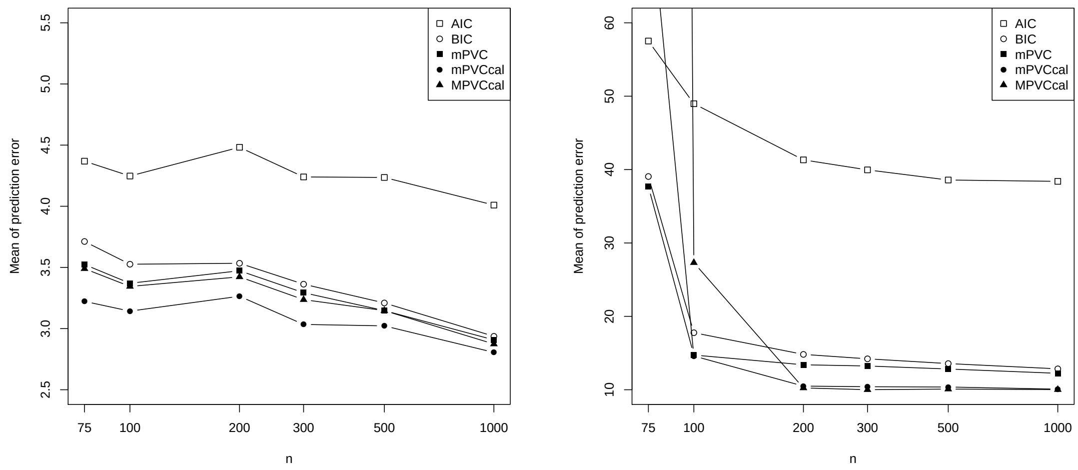


(c) (d)

Figure 1: Estimated probabilities of correct model selection for models M1 (a), M2 (b), M3 (c) and M4 (d) with respect to n (on a logarithmic scale) based on $N = 500$ trials.

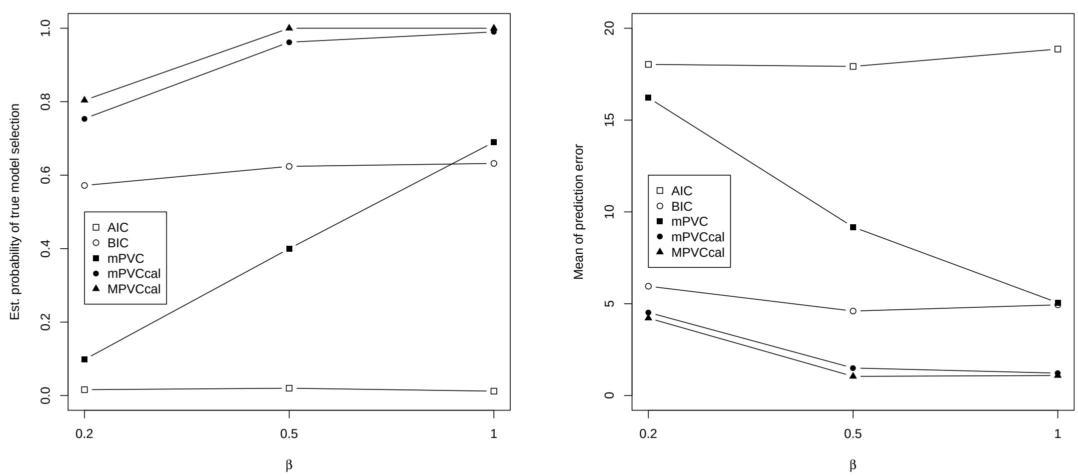


(a) (b)



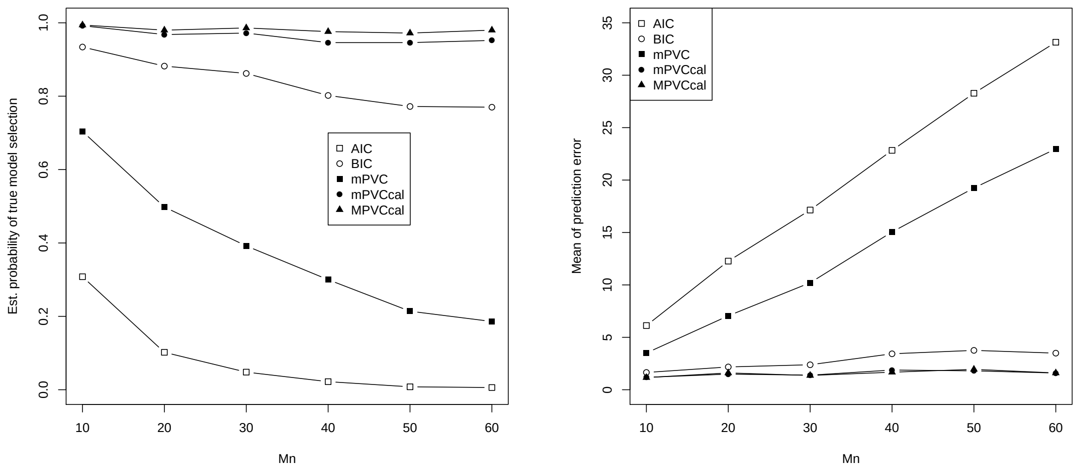
(c) (d)

Figure 2: Means od prediction error for models M1 (a), M2 (b), M3 (c) and M4 (d) with respect to n (on a logarithmic scale) based on $N = 500$ trials.



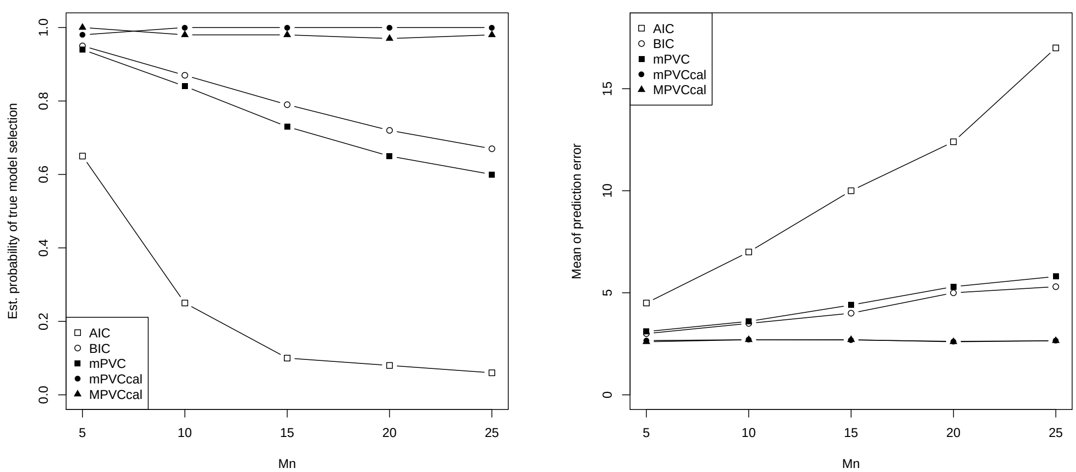
(a) (b)

Figure 3: Estimated probabilities of correct model selection (a) and means of prediction error (b) with respect to value of parameter β for model M1 for sample size $n = 300$ based on $N = 500$ trials.



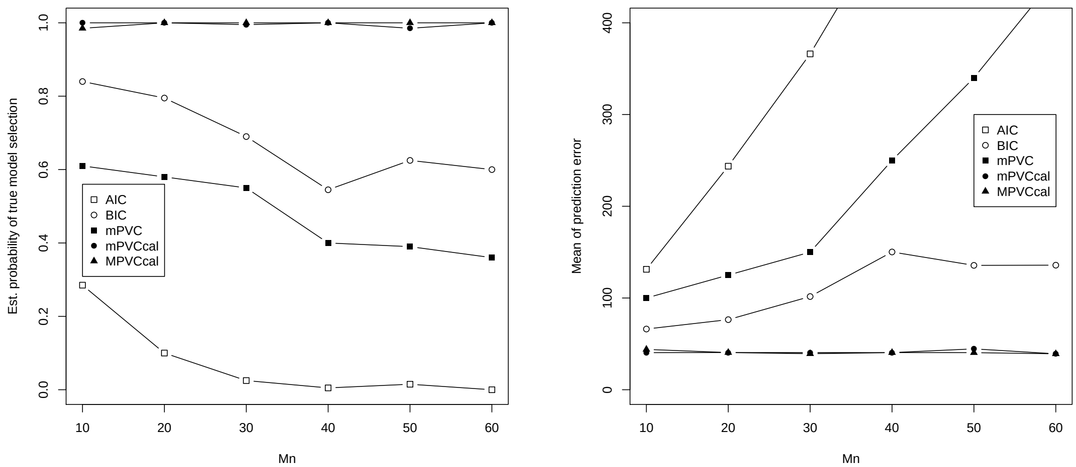
(a) (b)

Figure 4: Estimated probabilities of correct model selection (a) and means of prediction error (b) with respect to M_n for model M1 for sample size $n = 1000$ based on $N = 500$ trials.



(a) (b)

Figure 5: Estimated probabilities of correct model selection (a) and means of prediction error (b) with respect to M_n for model (L1) based on $N = 500$ trials.



(a) (b)

Figure 6: Estimated probabilities of correct model selection (a) and means of prediction error (b) with respect to M_n for `bodyfat` data set.

6 Appendix

Proof of Lemma 1

The lemma is proved in Pokarowski and Mielniczuk (2010). For completeness we give an outline of proof here. Recall that $B_{a,b}$ and $B(x, y)$ denote a random variable having beta distribution with shape parameters a and b and beta function, respectively. Let $B_x(a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$ be the incomplete beta function. It can be easily proved that

$$aB_x(a, b) = x^a(1-x)^b + (a+b)B_x(a+1, b), \tag{20}$$

and

$$B_{1-x}(b, a) = B(a, b) - B_x(a, b). \tag{21}$$

Consider the case $a \geq 1$. Using (20), (21) and assumption $x > \frac{a-1}{a+b}$ we obtain the upper bound in (6)

$$P[B_{a,b} > x] = 1 - \frac{B_x(a,b)}{B(a,b)} = \frac{B_{1-x}(b,a)}{B(a,b)} =$$
$$\frac{1}{B(a,b)b} \cdot (1-x)^b x^a [1 + \frac{a+b}{b+1}(1-x) + \frac{(a+b)(a+b+1)}{(b+1)(b+2)}(1-x)^2 + \ldots] \leq$$
$$\frac{1}{B(a,b)b} \cdot (1-x)^b x^a [1 + \frac{a+b}{b+1}(1-x) + \left(\frac{a+b}{b+1}\right)^2 (1-x)^2 + \ldots] = \frac{(1-x)^b x^{a-1}}{B(a,b)b}(1 + L(a,b,x)).$$

In order to obtain the lower bound in (6) note that for $a \geq 1$

$$\frac{1}{B(a,b)b} \cdot (1-x)^b x^a [1 + \frac{a+b}{b+1}(1-x) + \frac{(a+b)(a+b+1)}{(b+1)(b+2)}(1-x)^2 + \ldots] \geq \frac{1}{B(a,b)b} \cdot (1-x)^b x^{a-1}.$$

The case $a < 1$ can be treated analogously.

For ease of notation we assume in the following proofs that $\sigma^2 = 1$. Let $\mathbf{Q}(j)$ denote projection on the column space spanned by the regressors corresponding to coefficients in a given model j.

Proof of Lemma 3

Consider first the case $j \subset t$. Denote $\mathbf{W} = \mathbf{E}(\mathbf{x}_{lt}\mathbf{x}_{lt}')$, which in view of assumption (A0) is positive definite. Define $\Lambda_{n,j} = n^{-1}(\mathbf{X}\boldsymbol{\beta})'[\mathbf{I} - \mathbf{Q}(j)](\mathbf{X}\boldsymbol{\beta}) > 0$. Let $\mathbf{D}_j$ be a $M_n \times j$ matrix of zeros and ones such that $\mathbf{X}\mathbf{D}_j$ consists of only these j columns of $\mathbf{X}$ which correspond to model j. By assumption (A0) and using the fact that $\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}\mathbf{D}_t)\bar{\boldsymbol{\beta}}$ where $\bar{\boldsymbol{\beta}} = (\beta_{t_1}, \ldots, \beta_{t_{p_t}})'$ we have $\Lambda_{n,j} \xrightarrow{P} \lambda > 0$ as $n \to \infty$. The assertion follows from the fact that for $j \subset t$

$$n^{-1}(\mathbf{X}\boldsymbol{\beta})'[\mathbf{I} - \mathbf{Q}(j)](\mathbf{X}\boldsymbol{\beta}) = n^{-1}\bar{\boldsymbol{\beta}}'\mathbf{A}\bar{\boldsymbol{\beta}}, \tag{22}$$

where

$$\mathbf{A} = [(\mathbf{X}\mathbf{D}_t)'(\mathbf{X}\mathbf{D}_t)] - [(\mathbf{X}\mathbf{D}_t)'(\mathbf{X}\mathbf{D}_t)]\bar{\mathbf{D}}_j[\bar{\mathbf{D}}_j'(\mathbf{X}\mathbf{D}_t)'(\mathbf{X}\mathbf{D}_t)\bar{\mathbf{D}}_j]^{-1}\bar{\mathbf{D}}_j'[(\mathbf{X}\mathbf{D}_t)'(\mathbf{X}\mathbf{D}_t)]$$

and $\bar{\mathbf{D}}_j$ is a $p_t \times p_j$ matrix such that $\mathbf{X}\mathbf{D}_j = (\mathbf{X}\mathbf{D}_t)\bar{\mathbf{D}}_j$. Matrix $\mathbf{W}$ as a positive definite matrix can be decomposed as $\mathbf{W} = \mathbf{W}^{1/2}\mathbf{W}^{1/2}$ where $\mathbf{W}^{1/2} = \mathbf{U}\Xi^{1/2}\mathbf{U}'$, $\mathbf{U}$ is an orthogonal matrix and Ξ is a diagonal matrix with positive diagonal. The right hand side of (22) converges in probability to

$$\begin{aligned} &\lambda = \bar{\boldsymbol{\beta}}'[\mathbf{W} - \mathbf{W}\bar{\mathbf{D}}_j(\bar{\mathbf{D}}_j'\mathbf{W}\bar{\mathbf{D}}_j)^{-1}\bar{\mathbf{D}}_j'\mathbf{W}]\bar{\boldsymbol{\beta}} = \\ &(\mathbf{W}^{1/2}\bar{\boldsymbol{\beta}})'[\mathbf{I} - \mathbf{W}^{1/2}\bar{\mathbf{D}}_j(\bar{\mathbf{D}}_j'\mathbf{W}\bar{\mathbf{D}}_j)^{-1}\bar{\mathbf{D}}_j'(\mathbf{W}^{1/2})']\mathbf{W}^{1/2}\bar{\boldsymbol{\beta}} > 0 \end{aligned}$$

since the columns of $\mathbf{W}^{1/2}$ are linearly independent. We have the following decomposition for $j \subset t$

$$n^{-1}RSS(j) = n^{-1}\boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{Q}(j))\boldsymbol{\varepsilon} + n^{-1}2(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{Q}(j))\boldsymbol{\varepsilon} + \Lambda_{n,j}. \tag{23}$$

The first summand converges in probability to σ^2. The last summand $\Lambda_{n,j} \xrightarrow{P} \lambda > 0$, as has been already shown. Provided that $\mathbf{X}'\mathbf{X}$ is invertible, $n^{-1}2(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{Q}(j))\boldsymbol{\varepsilon}$ given $\mathbf{X}$ has $N(0, v_n)$ distribution, where $v_n = n^{-1}\Lambda_{n,j} \xrightarrow{P} 0$. Thus $n^{-1}2(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{Q}(j))\boldsymbol{\varepsilon} \xrightarrow{P} 0$. This completes the first part of the proof. For $j \supseteq t$ the second and the third term in (23) are equal to zero. This yields the second part of the assertion.

Proof of Lemma 4

Define $b_n = n(\exp(R_n/n) - 1)$. It is easily seen that $b_n \geq R_n$ thus b_n satisfies the condition imposed on R_n. For $M_n = p_t$ the assertion is obvious, thus we assume that $M_n > p_t$

We have the following inequality

$$P\left\{n \log\left[\frac{RSS(t)}{RSS(f)}\right] > R_n\right\} = P\left\{\frac{RSS(t)}{RSS(f)} > \exp\left(\frac{R_n}{n}\right)\right\} =$$

$$
\begin{aligned}
&P\{\boldsymbol{\varepsilon}'[\mathbf{Q}(f)-\mathbf{Q}(t)]\boldsymbol{\varepsilon} > b_n n^{-1}\boldsymbol{\varepsilon}'[\mathbf{I}-\mathbf{Q}(f)]\boldsymbol{\varepsilon}\} \leq \\
&P\{\boldsymbol{\varepsilon}'[\mathbf{Q}(f)-\mathbf{Q}(t)]\boldsymbol{\varepsilon} > b_n n^{-1}(n-M_n-d_n)\}+ \\
&P\{\boldsymbol{\varepsilon}'[\mathbf{I}-\mathbf{Q}(f)]\boldsymbol{\varepsilon} \leq n-M_n-d_n\},
\end{aligned}
$$

where $d_n = (n-M_n)^{(1+\delta)/2}$, for some $\delta \in (0,1)$. Matrix $\mathbf{X}'\mathbf{X}$ has rank M_n and it follows that $\boldsymbol{\varepsilon}'[\mathbf{Q}(f)-\mathbf{Q}(t)]\boldsymbol{\varepsilon} \sim \chi^2_{M_n-p_t}$ and $\boldsymbol{\varepsilon}'[\mathbf{I}-\mathbf{Q}(f)]\boldsymbol{\varepsilon} \sim \chi^2{}_{n-M_n}$ (since $\sigma^2 = 1$). By an inequality for cumulative distribution function of a chi-square distribution,

$$
P(\chi^2_k \leq k-\delta_0) \leq \exp\{-(4k)^{-1}\delta_0^2\},
$$

for $\delta_0 > 0$ (see Shibata (1981)). Thus we have

$$
P\{\boldsymbol{\varepsilon}'[\mathbf{I}-\mathbf{Q}(f)]\boldsymbol{\varepsilon} \leq n-M_n-d_n\} \leq \exp\left[-\frac{d_n^2}{4(n-M_n)}\right] \to 0,
$$

as $n \to \infty$, since $M_n/n \to 0$. Let $\gamma_n = b_n(1-M_n/n-d_n/n)$. As $\boldsymbol{\varepsilon}'[\mathbf{Q}(f)-\mathbf{Q}(t)]\boldsymbol{\varepsilon} \sim \chi^2_{M_n-p_t}$ by Chebyschev inequality we have

$$
P\{\boldsymbol{\varepsilon}'[\mathbf{Q}(f)-\mathbf{Q}(t)]\boldsymbol{\varepsilon} - (M_n-p_t) > \gamma_n - (M_n-p_t)\} \leq \frac{2(M_n-p_t)}{[\gamma_n-(M_n-p_t)]^2} \to 0,
$$

where the last convergence follows from $(\gamma_n - M_n)/\sqrt{M_n} \to \infty$. This completes the proof.

Proof of Lemma 5

In view of conditions (A1.3) and (A1.4) matrix $(\mathbf{X}'\mathbf{X})^{-1}$ exists with probability tending to one (see the proof of Theorem 2 in Zheng and Loh (1997)). Recall that T_k is a t-statistic corresponding to the kth variable. It suffices to prove that for any $c_n \to 0$ $P[\min_{i\in t}\log(RSS(f-\{i\})/RSS(f)) < c_n] \to 0$. Noting that

$$
\frac{RSS(f-\{i\})}{RSS(f)} = \frac{T_i^2}{n-M_n} + 1,
$$

we obtain that

$$
\begin{aligned}
&P[\min_{i\in t}\log\frac{RSS(f-\{i\})}{RSS(f)} < c_n] \leq P[\min_{i\in t}T_i^2 < (n-M_n)(\exp(c_n)-1)] \\
&\leq P(\min_{i\in t}T_i^2 < (n-M_n)(\exp(c_n)-1)).
\end{aligned}
$$

Since $\exp(c_n)-1 = c_n + o(c_n)$ it suffices to show that $P[\min_{i\in t}T_i^2 < Cnc_n] \to 0$, for some $C > 0$. This follows from the proof of Theorem 2 in Zheng and Loh (1997) who proved that under conditions of this Lemma $P[\min_{i\in t}\hat{\sigma}^2 T_i^2 < nc_n] \to 0$, for any c_n such that $c_n \to 0$. Now the required convergence follows from the fact that $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$.

References

H. Akaike. Statistical predictor identification. *Annals of the Institute for Statistical Mathematics*, 22: 203–217, 1970.

G. Casella, J. Giron, M.L. Martinez, and E. Moreno. Consistency of bayes procedures for variable selection. *Annals of Statistics*, 37:1207–1228, 2009.

A. Davison and D. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, 1997.

R. W. Johnson. Fitting percentage of body fat to simple body measurements. *Journal of Statistical Education*, 4(1), 1996.

T. Ledwina. Data driven version of the Neyman smooth test of fit. *J. Amer. Statist. Assoc.*, 89:1000–1005, 1994.

E. Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 21:255–285, 1993.

E. Moreno, F. J. Giron, and G. Casella. Consistency of objective bayes factors as the model dimension grows. *Annals of Statistics*, 38:1937–1952, 2010.

P. Pokarowski and J. Mielniczuk. Linear model selection using p-values. *in preparation*, 2010.

B. Pötscher, M. and H. Leeb. Model selection. *Handbook of Financial Time Series*, pages 889–926, 2008.

A. C. Rencher and G. B. Schaalje. *Linear Models in Statistics*. Wiley, New Jersey, 2008.

J. Shao. An asymptotic theory for linear model selection (with discussion). *Statistica Sinica*, 7:221–264, 1997.

R. Shibata. An optimal selection of regression variables. *Biometrika*, 63:117–126, 1981.

X. Zheng and W-Y. Loh. A consistent variable selection criterion for linear models with high-dimensional covariates. *Statistica Sinica*, 7:311–325, 1997.